

Mixture models applied to heterogeneous populations

Carolina Valani Cavalcante* and Kelly Cristina M. Gonçalves†

Abstract

Mixture models provide a flexible representation of heterogeneity in a finite number of latent classes. From the Bayesian point of view, Markov Chain Monte Carlo methods provide a current way to draw inference from these models. In particular, when the number of subpopulations is considered unknown, more sophisticated methods are required to perform the Bayesian analysis. The Reversible Jump Markov Chain Monte Carlo is an alternative method for computing the posterior distribution by simulation in this case. Some problems associated with the Bayesian analysis of these class of models are frequent, as the so-called “label-switching”. However, as the level of the heterogeneity in the population increases, it is expected that these problems become less frequent and the model’s performance improves. Thus, the aim of this work is to evaluate the normal mixture model fit using simulated data under different settings of heterogeneity and prior information about the mixture proportions. A simulation study was also carried out to evaluate the model’s performance considering the number of components known and estimating it. Finally, the model is applied to a real data set that consists of antibody levels of Cytomegalovirus in individuals.

Keywords: identifiability, sensitivity analysis, subpopulations, frequentist properties, NHANES

*Escola Nacional de Ciências Estatísticas (ENCE), Rio de Janeiro, RJ, Brazil. Email: carolinavalanic@gmail.com

†Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brazil. Email: kelly@dme.ufrj.br

1 Introduction

Mixture models are noted for their flexibility in modeling complex data and are widely used in the statistical literature (see [6]). Such models provide a natural framework for the modeling of heterogeneity in a population. Moreover, due to the large class of functions that can be approximated by mixture models, they are attractive for describing non-standard distributions and have been adopted in many areas, as genetics, ecology, computer science, economics, biostatistics and many others. For instance, as stated in [4], in genetics, location of quantitative traits on a chromosome and interpretation of microarrays are both related to mixtures, while, in computer science, spam filters and web context analysis start from a mixture assumption to distinguish spams from regular emails and group pages by topic, respectively.

Statistical analysis of mixtures has not been straightforward and the Bayesian paradigm has been particularly suited to their analysis. This framework allows the complicated structure of a mixture model to be decomposed into a set of simpler structures through the use of hidden or latent variables. According to [8], when the number of components is unknown, the Bayesian paradigm is the only sensible approach to its estimation. Then, the Bayesian approach contributed to mixture models become increasingly popular in many areas. In real applications, the number of components can arise important conclusions about the problem, so it has to be well specified or estimated, despite we usually have little theoretical guidance. On the other hand, even if prior theory suggests a particular number of components we may not be able to reliably distinguish between some of the components. In some cases additional components may simply reflect the presence of outliers in the data.

When the number of subpopulations is assumed known, Markov Chain Monte Carlo methods (MCMC) can be used for Bayesian estimation of the subpopulation parameters. Nevertheless, this method, as originally formulated, requires the posterior distribution to have a density with respect to some fixed measure. When the number of components is considered unknown, i.e., the size of the parametric space is also a parameter, it appears a problem with variable dimension, thus MCMC cannot be used alone in this case and more sophisticated methods are required to perform the Bayesian analysis. One alternative in this case is the approach based on Reversible Jump MCMC (RJMCMC),

which was first proposed in [2] and applied in univariate Normal mixture models with unknown numbers of components by [8]. The method basically consists of jumps between the parameter subspaces corresponding to different numbers of components in the mixture.

Whilst MCMC provides a convenient way to draw inference from complicated statistical models, there are still many, perhaps under appreciated, problems associated with the MCMC analysis of mixtures. The problems are mainly caused by the nonidentifiability of the components under symmetric priors, which leads to the so called label-switching in the MCMC output, discussed in [3]. The term describes the invariance of the likelihood under relabelling of the mixture components, which can lead to the posterior distribution of the parameters being highly symmetric and multimodal, making it difficult to summarize. In particular, the usual practice of estimating parameters by their posterior mean, and summarizing joint posterior distributions by marginal distributions is often inappropriate. One frequent response to this problem is to remove the symmetry by using artificial identifiability constraints. This and other alternative classes of approaches to this problem are described by [10].

The aim of this work is to review and discuss the application of mixture models, in particular the Normal mixture models, to heterogeneous populations under the Bayesian approach. The main purpose is to evaluate the model's performance in different settings of heterogeneity and considering the number of components known and unknown. We verify if the label-switching phenomena generally persists when the subpopulations are not well separated, then, if the more the population is heterogeneous, the more the model parameters should be better estimated. Furthermore, we evaluate the label-switching assuming more informative prior distributions for the mixture weights.

The paper is organized as follows. Section 2 presents the general definition of a mixture model and discusses some aspects of the inference. A simulation study for assessing the estimation of model parameters under different levels of heterogeneity is presented in Section 3. Additionally, a prior sensitivity analysis of the mixture proportions is presented. It also discusses the model fit when the number of components is known and unknown. In Section 4 the performance of the methodology is assessed

through an application to a real data set. Finally, Section 5 presents some conclusions and suggestions for further research.

2 Finite mixture models

The basic mixture model for independent scalar or vector observations Y_i , $i = 1, \dots, n$, is a convex combination given by:

$$Y_i \sim \sum_{j=1}^k w_j f(\cdot \mid \boldsymbol{\theta}_j), \quad i = 1, \dots, n, \quad (1)$$

where $f(\cdot \mid \boldsymbol{\theta})$ is a given parametric family of densities indexed by a scalar or a vector $\boldsymbol{\theta}$. In general, the objective of the analysis is to make inferences about the unknowns: the number of components, k ; the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k)$ with $\boldsymbol{\theta}_j$ being specific to component j ; and the components' weights, $\mathbf{w} = (w_1, \dots, w_k)$, $0 < w_j < 1$, $\sum_{j=1}^k w_j = 1$. Let $\boldsymbol{\Phi} = (\mathbf{w}, \boldsymbol{\theta}, k)$ be the parametric vector of the model (1).

For a random sample $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ observed, the likelihood function of $\boldsymbol{\Phi}$ is given by:

$$p(\mathbf{y} \mid \boldsymbol{\Phi}) = \prod_{i=1}^n \sum_{j=1}^k w_j f(y_i \mid \boldsymbol{\theta}_j).$$

The likelihood function leads to k^n terms, what brings a computational difficulty.

A context in which the model (1) can arise and we are interested in this paper is when we postulate a heterogeneous population consisting of heterogeneous groups $j = 1, 2, \dots, k$ of sizes proportional to w_j , from which a random sample is drawn. The label of the group from which each observation is drawn is unknown and it is natural to regard the group label z_i , for the i -th observation as a latent variable and rewrite (1) as the following hierarchical model: for $i = 1, \dots, n$, $j = 1, \dots, k$,

$$Y_i \mid \boldsymbol{\theta}_j, z_i = j \sim f(\cdot \mid \boldsymbol{\theta}_j), \quad \text{with } P(z_i = j) = w_j. \quad (2)$$

Integrating $\mathbf{z} = (z_1, \dots, z_k)$ out from (2) we return to model (1). The formulation given by (2) is convenient for interpretation and calculation, decreasing the computational cost.

A Bayesian approach to inference requires the specification of a prior distribution $p(\cdot)$ for the parameters of the mixture model (1). In particular, the prior elicitation is an important question. Being fully non-informative and obtaining proper posterior distributions are not possible in a mixture context. An alternative on this case is to define weakly informative priors, which may or may not be data dependent.

The mixture model in (2) is invariant to permutation of the labels $j = 1, \dots, k$. Some implications of this for likelihood analyses are discussed by [7]. If we have no prior information that distinguishes between the components of the mixture, so the prior distribution $p(\cdot)$ is the same for all permutations of $\boldsymbol{\theta}$, then the posterior distribution will be similarly symmetric and, there will be $k!$ symmetric modes of the posterior distribution. Therefore, the component labels are mixed up and cannot be distinguished from each other. As a result, the marginal on the parameters for all components is identical and the posterior expectation for the parameters is identical too, and so estimating the parameters on the basis of the MCMC output is not straightforward.

There are some suggested solutions to this problem, see [10] for details. One common response to the label-switching problem is to impose an identifiability constraint on the parameter space. This breaks the symmetry of the prior and thus, of the posterior distribution of the parametric vector. For example, we can impose an ordering constraint on θ_j 's, such as $\theta_1 < \theta_2 < \dots < \theta_k$, if it is a scalar. However, for any given data set, many choices of identifiability constraint may be ineffective in removing the symmetry in the posterior distribution.

2.1 Inference

As we are in a Bayesian framework, the inference consists in obtain the posterior distribution of the parametric vector $\boldsymbol{\Phi}$ of model (2). In general, this distribution cannot be obtained in closed form. Therefore, it is necessary to use some numerical approximation methods. One alternative, which is often used and is feasible to implement, is to generate samples from the marginal distributions of the parameters

based on the MCMC algorithm. A comprehensive Bayesian treatment using MCMC methods was presented in [1] for finite mixture models.

Nevertheless, this method, as originally formulated, requires the posterior distribution to have a density with respect to some fixed measure. Thus, in the mixture context, the method can only be applied when the number of components k in the model (2) is considered known.

However, rarely the number k is known, and fix it on an incorrect value can bring important consequences to the posterior distribution. Other times, the target of the study is exactly the estimation of k . The approach based on RJMCMC is an alternative in this case, which was proposed on this context by [8]. It operates on the augmented parameter space, where the allocation variables \mathbf{z} are included as unknown parameters.

The method basically consists of jumps between the parameter subspaces corresponding to different numbers of components in the mixture, after updating them. If the current model is a mixture with $k > 1$ components, then it is usual to reduce the searching strategy to moves that either preserve the number of components, or lead to a mixture with $k - 1$ or $k + 1$ components. The idea is then to supplement each of the spaces with adequate artificial spaces in order to create a bijection between them, most often by augmenting the space of the smaller model. Jumps are achieved by adding new components, deleting existing components, and splitting or merging these. These moves are randomly chosen and after being drawn makes necessary corresponding changes to $(\boldsymbol{\theta}, \mathbf{w})$.

With respect to the label-switching problem, during MCMC computation, the sampler should switch from modes to modes between the iterations, however the failure to visit the identical posterior expectations reveals that the MCMC sampler has not converged and the posterior distribution surface is multimodal. If the value of $k!$ is very large, it would be hard for the regular MCMC sampler to thoroughly and explore the high multimodality. Thus, the MCMC samples might be trapped into local modes and it would require an enormous number of iterations to escape from it and the label switch would cause very poor estimates and the results might be very different from different runs of the MCMC.

2.2 Normal mixture model

In this work we are particularly interested in the univariate Normal case presented in [8], then θ_j in (1) becomes a vector with expectation and variance parameters (μ_j, σ_j^2) . The model is stated below: for $i = 1, \dots, n$ and $j = 1, \dots, k$,

$$\begin{aligned} Y_i \mid \mu_j, \sigma_j^2, z_i = j &\sim \text{Normal}(\mu_j, \sigma_j^2), \\ P(z_i = j) &= w_j. \end{aligned} \tag{3}$$

Assuming that the parameters in Φ are prior independent and identically distributed and that k is unknown, the prior distribution is given by: for $i = 1, \dots, n$ and $j = 1, \dots, k$,

$$\begin{aligned} \mathbf{w} &\sim \text{Dirichlet}(\gamma), \\ \mu_j &\sim \text{Normal}(\mu_a, \sigma_a^2), j = 1, \dots, k, \\ \sigma_j^{-2} &\sim \text{Gamma}(\alpha, \beta), j = 1, \dots, k, \\ \beta &\sim \text{Gamma}(g, h), \\ k &\sim \text{Uniform}\{1, k_{\max}\}, \end{aligned} \tag{4}$$

where $\text{Dirichlet}(a)$ generically denotes the symmetric Dirichlet distribution with parameter a . The symmetric Dirichlet distributions are often used, since there typically is no prior knowledge favoring one component over another. Since all elements of the parameter vector have the same value, the distribution alternatively is parametrized by a single scalar value a . $\text{Gamma}(a, b)$ represents the gamma distribution with mean a/b and variance a/b^2 and $\text{Uniform}\{a, \dots, b\}$ is the Uniform distribution defined on the integers $\{a, \dots, b\}$. Moreover, for identifiability, we can use for example that the μ_j are in increasing numerical order, thus the joint prior distribution of Φ is $k!$ times the product of their marginal prior distributions.

In this paper, as treated in [8] we considered the Bayesian estimation in the set-up where we do not have strong prior information on the mixture parameters. On the other hand, being fully non-informative and obtaining proper posterior distributions are not possible in a mixture content. An alternative on this case is to keep the simple independence and define weakly informative priors, which may or may not be data

dependent. Therefore, the default hyperparameter choices can be viewed with further details in [8].

Furthermore, in a normal mixture model the posterior distribution of the means for example could overlap, but the extent of the overlap depends on its separation and the sample size. When the means are well separated, labels of the realizations from the posterior by ordering their means generally coincide with the population ones. As the separation reduces, label-switching may occur. This problem can be also minimized by choosing to order other parameters of the mixture components, for example, the variance, weights or some combination of all three parameters.

In this work the unique labeling will be achieved by imposing a restriction on μ_j . We will use that in which the μ_j are in increasing numerical order; thus the joint prior distribution of the parameters is $k!$ times the prior density, restricted to the set $\mu_1 < \mu_2 < \dots < \mu_k$.

3 Simulation study

To assess the convergence of the MCMC and RJMCMC estimation, we generated some samples under different settings. We obtained samples from the posterior distributions of the model parameters, supposing k known and estimating it. The population estimates were then compared with the true values to evaluate the model's performance. The aim is to evaluate the performance of the Normal mixture model varying the level of heterogeneity and the prior information elicited for the mixture proportions. Furthermore, we also compared the results obtained under each simulation method considered, MCMC and RJMCMC.

3.1 Assessment of RJMCMC and MCMC under different scenarios

To check the convergence of the RJMCMC and MCMC estimations, we generated one sample with $n = 100$ observations under two levels of heterogeneity, the first one with groups well separated, which we call as the more heterogeneous sample and the other with groups less well separated, which represents the more homogeneous one. On both

scenarios we fixed $k = 5$, $\sigma^2 = (\sigma_1^2, \dots, \sigma_5^2) = (0.22, 1.95, 0.92, 0.74, 1.13)$ and $\mathbf{w} = (w_1, \dots, w_5) = (0.17, 0.21, 0.34, 0.12, 0.16)$. With this value fixed for \mathbf{w} we expected to have groups with a reasonable number of observations, thus we did not consider scenarios with groups outliers. The heterogeneous scenario was obtained fixing $\boldsymbol{\mu} = (\mu_1, \dots, \mu_5) = (-3, 0, 4, 11, 16)$ and the homogeneous fixing $\boldsymbol{\mu} = (\mu_1, \dots, \mu_5) = (0, 2, 4, 6, 8)$. Figure 1 presents the distribution of both data generated. The aim of this study is to verify if the level of heterogeneity of the population affects the results, mainly the label-switching problem.

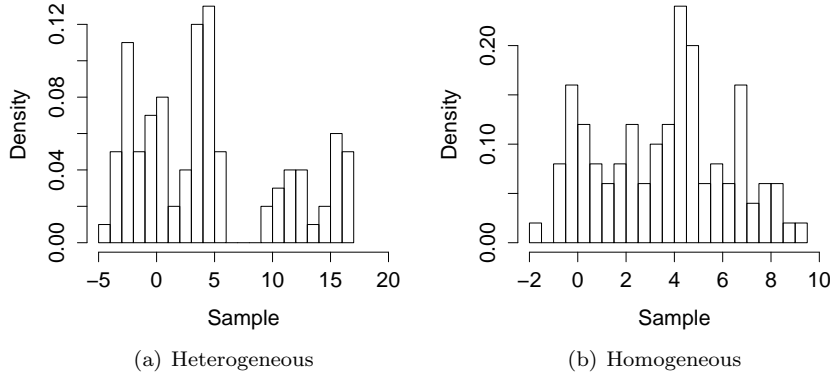


Figure 1: *Histograms with the distribution of the samples generated.*

The prior distribution considered are described in (4), and we elicited the prior for μ_j and σ_j^2 using the same idea of weakly informative prior suggested by [8]. First, we assumed k unknown and in its prior distribution presented in (4) we assumed $k_{max} = 10$, thus RJMCMC was used to obtain samples from the posterior distribution. We also did a brief prior sensitivity analysis assuming two values on the Dirichlet prior distribution for each data: $\gamma \in \{1, 4\}$ for the heterogeneous case and $\gamma \in \{1, 4, 10\}$ for the homogeneous case. To assume $\gamma = 1$ is equivalent to a uniform distribution over all points in its support. On the other hand, the parameter value above 1 gives some information that all sample proportions in subpopulations are similar to each other.

For the RJMCMC simulations, we generated, respectively for the homogeneous sample and the heterogeneous one, 350,000 and 70,000 samples from the posterior distribution, discarded the first 10,000 and 20,000, and then thinned the chain by taking every 10th sample value. Figure 2 displays the histogram with the posterior densities of

k for some values of γ . It should be noted that for the heterogeneous case the parameter k is well estimated, but when $\gamma = 4$ the estimate is more accurate. On the other hand, k is underestimated when assuming $\gamma = 4$ with the homogeneous sample. The same happen with $\gamma = 1$ or any value less than 4. In this case, when $\gamma = 10$ the value of k is well estimated.

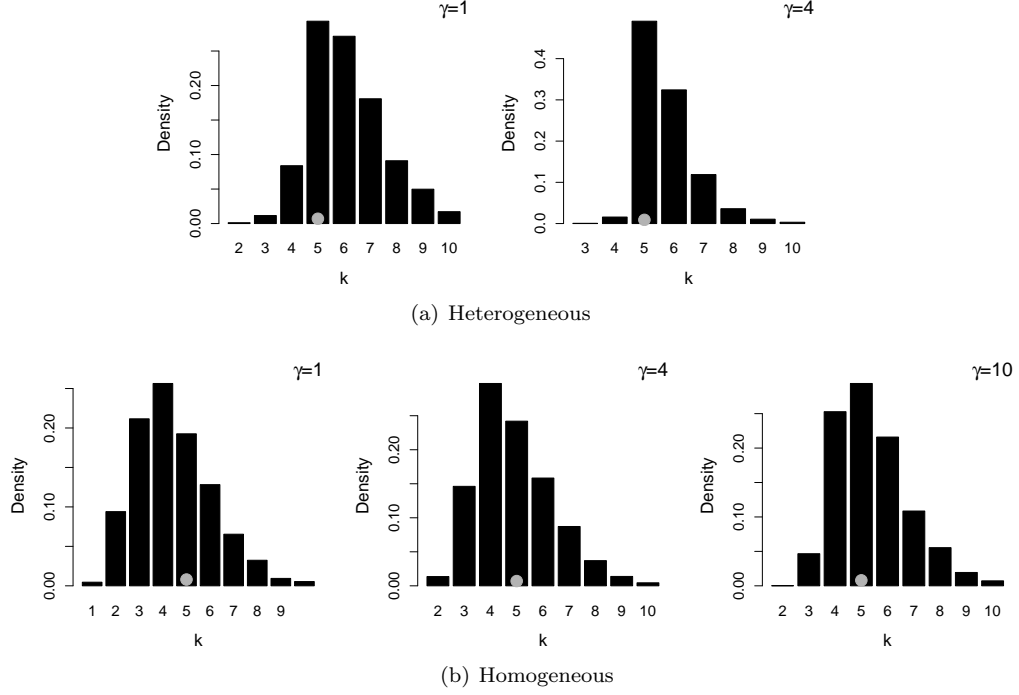


Figure 2: *Posterior densities of the parameter k for each values of γ considered in the prior distribution of the mixture proportions. The gray point represents the true value fixed in the simulation.*

Figure 3 shows the trace plot with the posterior distribution of parameters μ_j conditional on the posterior samples, whose estimated value of k is the one with higher posterior probability. Here, we also considered the value of k known and fixed it on the true value used to generate the samples, so MCMC was also used to generate samples from the posterior distribution. For the MCMC simulations, we generated 70,000 samples from the posterior distribution, discarded the first 20,000, and then thinned the chain by taking every 10th sample value, for both data generated.

All the results were obtained for each scenario and value of γ considered. The black trace represents the posterior density fixing $\gamma = 1$, the blue trace when $\gamma = 4$ in both scenarios and $\gamma = 10$ is represented by the red trace in the homogeneous case. The gray

line represents the true value of each μ_j . Note that on the homogeneous case, when RJMCMC is used, there is only a red trace for μ_5 , that is because the posterior for k favors the value 5, only when $\gamma = 10$. When analyzing Figure 3 it can be seen the effects of label-switching in the sampled values of the component means for many cases, even in the heterogeneous case. However this behavior improves when giving some prior information about the mixture proportions. It is also possible to observe that on the homogeneous case to obtain better results it is needed to increase even more the value of γ , this is, to give more prior information that the proportion observed in groups are similar.

The results obtained indicate that the RJMCMC and MCMC chains have converged for some cases, but for others the label-switching phenomena appears significantly, and so estimating the means on the basis of the RJMCMC and MCMC output is not straightforward. However, as the value of γ increases this behavior improves. If the number of iterations increases and then, the lag of the chain, the convergence may also improve, but it would require a high computational effort, thus we suggest here the carefully elicitation of the prior distribution to have better estimates. Almost all the mean parameters are well estimated when $\gamma = 4$ and $\gamma = 10$ in the heterogeneous and homogeneous case, respectively. Traces and density estimates for the mixture proportions and variances present this same behavior.

In general, MCMC and RJMCMC present a similar behavior, mainly for the heterogeneous sample generated. A more interesting comparison between both approaches is presented in the next subsection.

Figure 4 shows summary statistics of the posterior distributions of the mean parameters after reaching the supposed convergence for each of the scenarios and prior assumed, when assuming k unknown. The crosses represent the true value, the lines the 95% credibility interval and the points are the posterior mean. Also, the results in black are obtained assuming $\gamma = 1$, the blue one when $\gamma = 4$ and the red when $\gamma = 10$. In almost all the cases the intervals contain the true value. It is possible to observe the impact of the label-switching, which difficulties the parameters estimation, but also the improvement of the results when assuming a more informative prior to \mathbf{w} .

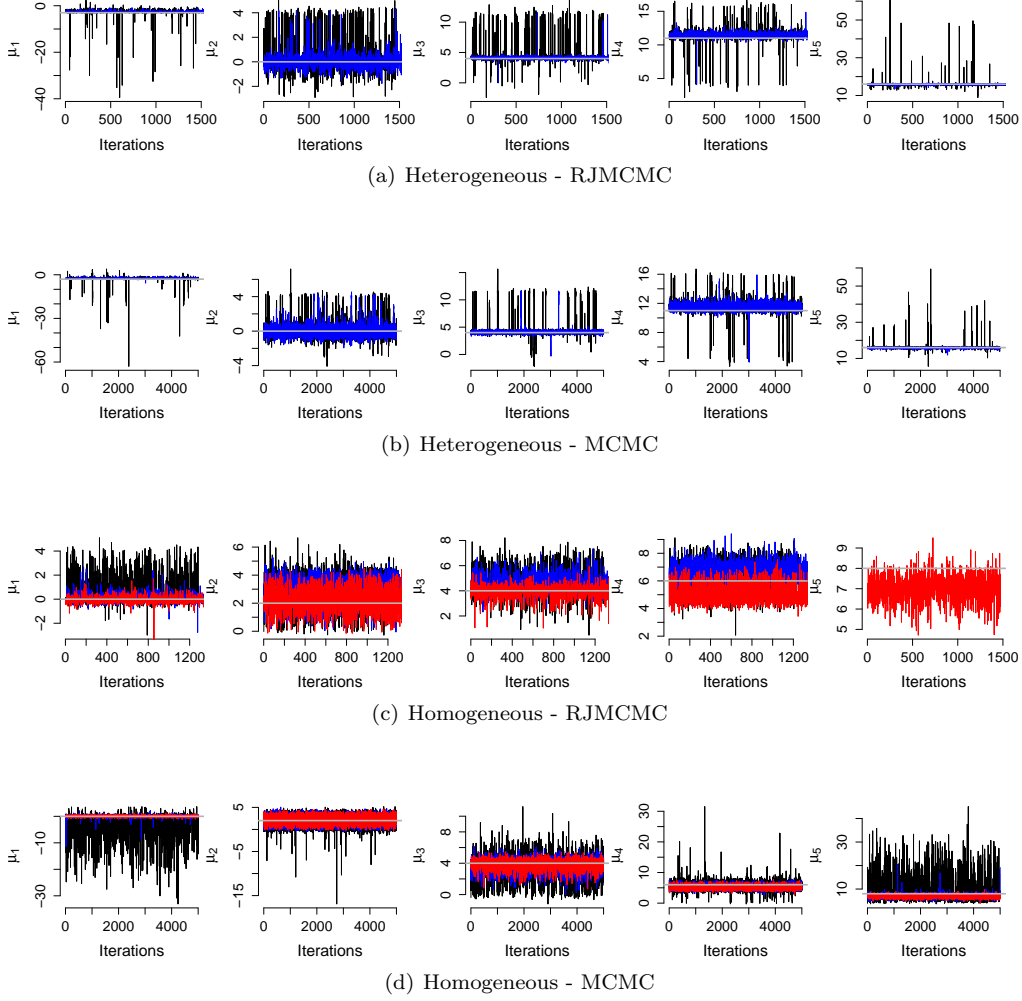


Figure 3: Trace plots with the posterior densities of the parameters of μ obtained from the fit of the normal mixture model under the different priors considered for \mathbf{w} and the two samples. We also assumed the value of k unknown (RJMCMC) and known (MCMC). The black trace is obtained assuming $\gamma = 1$, the blue one when $\gamma = 4$ and the red with $\gamma = 10$. The gray line represents the true value of each μ_j , $j = 1, \dots, 5$.

Therefore, we conclude that in those cases considered here the identifiability problem can be minimized under more informative priors and it was not necessary to use other alternative class of approaches to deal with the identifiability problem, as those described in [3]. The prior distribution of \mathbf{w} seems to have strong impact on the posterior distribution, improving the results, even in a homogeneous case. Furthermore, as the degree of heterogeneity increases, the mixture model performs considerably better even under less prior information. The same conclusions were attained when estimating the value of k or considering it known.

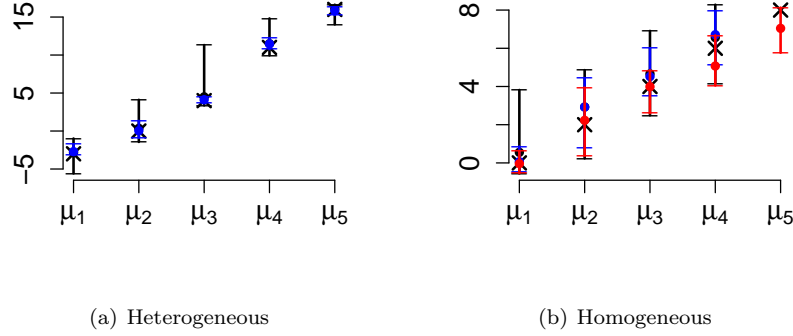


Figure 4: *Summary measurements for the point and 95% credibility interval estimates of the mean parameters model for a heterogeneous and homogeneous sample under three prior distributions for \mathbf{w} : the results in black are obtained assuming $\gamma = 1$, the blue one when $\gamma = 4$ and the red when $\gamma = 10$. Here it was considered the value of k unknown, so RJMCMC was used. The crosses represent the true value, the lines the 95% credibility interval and the points are the posterior mean.*

Additionally, Figure 5 shows the predictive densities for the two data sets generated, for all the prior distributions considered for \mathbf{w} , represented by the solid ($\gamma = 1$), dashed ($\gamma = 4$) and dotted ($\gamma = 10$) lines, respectively. The predictive densities in black are those obtained when the value of k is estimated, so RJMCMC was used, and the red ones are obtained when the value of k is fixed on the true value, so MCMC was used. The densities obtained under RJMCMC are conditional on the posterior samples whose sampled value of k is equal to the value with high posterior probability among all the samples. In contrast with the above results, an estimate of the predictive density based on the RJMCMC and MCMC outputs is unaffected by the label-switching problem, since it does not depend on how the components are labeled. It should be noted that the predictive density is better estimated on the heterogeneous sample than the homogeneous one and that the prior distribution does not affect the estimates. Moreover, the results obtained in the estimation considering k unknown and fixed are very similar to each other.

3.2 Comparison between RJMCMC and MCMC

To examine the performance of the Bayes estimators obtained under each simulation method, we generated two artificial samples of size $n = 100$, fixing k into two different

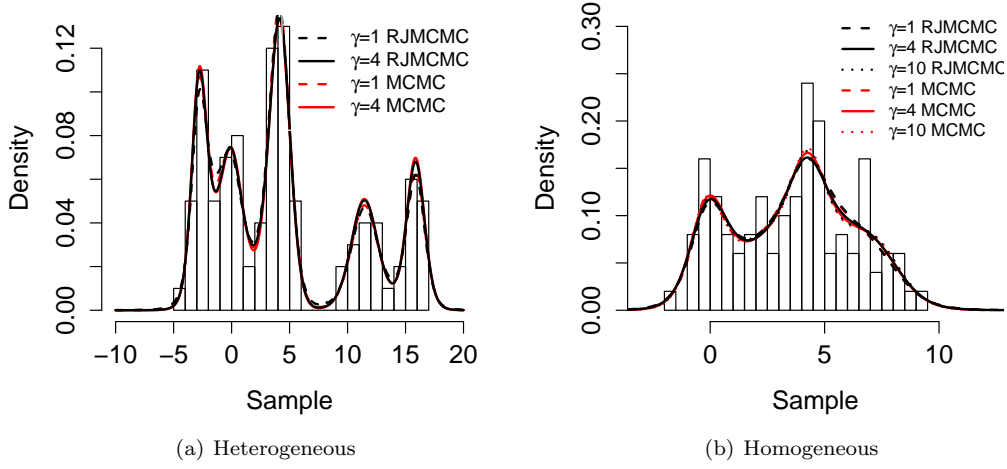
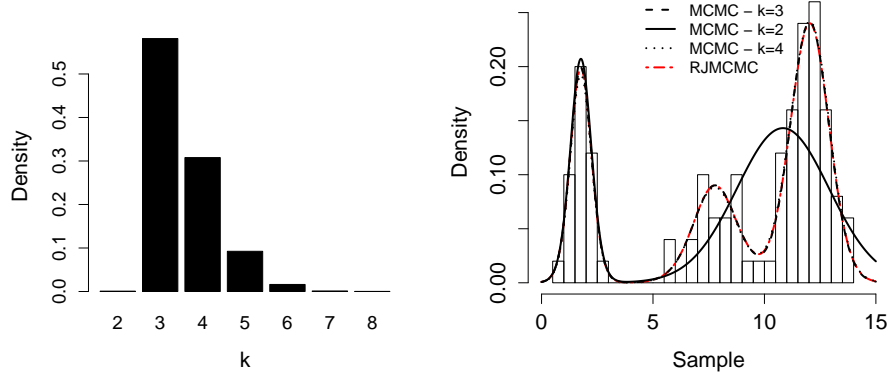


Figure 5: *Predictive densities considering different prior distributions for \mathbf{w} and estimating the value of k (RJMCMC) and fixing it on its true value (MCMC).*

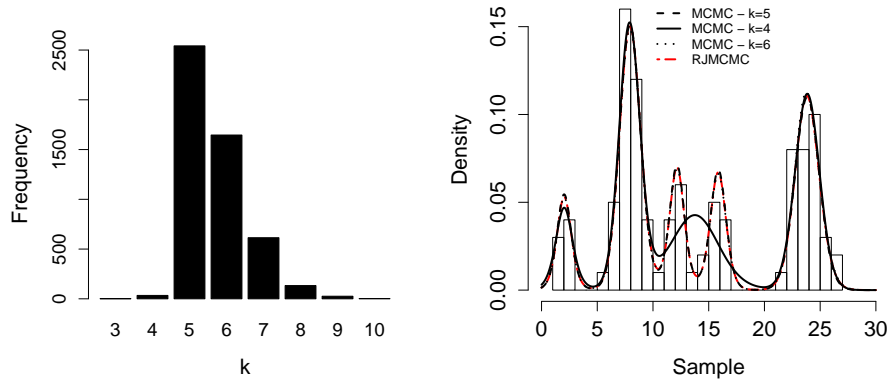
values, $k = 3$ and $k = 5$, in order to also evaluate the results when varying the value of k . Then, we obtained samples from the posterior distribution of the parametric vector, supposing k known (MCMC) and estimating it (RJMCMC). In the MCMC simulation we particularly fixed k for each case in three different values: we assumed it 2, 3 and 4 for the first sample and 4, 5 and 6 for the second one. We assumed here the same prior distribution used in Section 3.1. Thus, we are interested in evaluate the method's performance when we fix k on its true value, on a smaller and a greater value than the true one and when it is estimated. For the RJMCMC and MCMC simulations, we generated 350,000 samples from the posterior distribution, discarded the first 10,000, then thinned the chain by taking every 50th sample value, and the convergence was supposed achieved.

Figure 6 presents the posterior distribution of k obtained under the RJMCMC simulation and the predictive densities obtained for each sample generated. It should be noted that k is well estimated and all predictive densities are very similar, except when we fixed k in a value lower than the true one. Moreover, fixing k in a greater value than the truth does not affect the results.

Table 1 presents the Deviance Information Criterion (DIC), introduced by [9], for each approach considered in this study. DIC evaluates the goodness of fit of the model, thus the model with the smallest DIC should be the one that would best fit. The model



(a) $k = 3$



(b) $k = 5$

Figure 6: Predictive densities considering k known (MCMC) and fixing it in $k = 2, 3, 4$ and estimating it (RJMCMC) (RJMCMC).

with k known seems to fit the data better than its counterparts. However, the results are very similar, even when k is estimated, increasing the size of the parametric vector, except when k is fixed in a smaller value than the truth.

Table 1: DIC measurements for the models considered.

	MCMC ($k - 1$)	MCMC (k)	MCMC ($k + 1$)	RJMCMC
DIC ($k = 3$)	475.56	437.58	438.69	439.45
DIC ($k = 5$)	571.26	555.39	557.93	557.27

Thus, if the number of components is unknown and we use the MCMC algorithm to sample from the posterior distribution of the parametric vector, better results are attained fixing it equal or greater than the true value. On the other hand, to estimate the value of k and use the RJMCMC method is a good alternative in this case, having a similar performance to the case when we fix k in its true value.

Finally, we also generated 1,000 samples fixing the parameters on the previous values and obtained samples from the posterior distribution of the parametric vector, supposing k known and fixed on the true value in the MCMC and estimating it using the RJMCMC algorithm. The estimates were then compared with the true values to evaluate the model's performance.

First, in 89.9% of the 1,000 samples the value of k was correctly estimated when using RJCMC to sample from the posterior distribution. Table 2 shows summary statistics with some frequentist measures of the posterior distributions of the model parameters after reaching convergence. It reports the square root of the mean square error (SRMSE), the mean absolute error (MAE), the empirical nominal coverage of the 95% credibility intervals measured in percentages (Cov.) and the respective widths averaged over the 1,000 simulations (Wid.). In particular, the summary statistics of the components parameters are obtained conditioning on k in the value with higher posterior probability.

The parameters are well estimated in both cases and the results are very similar considering each approach, except the parameters σ_2^2 and σ_3^2 , which was slightly better estimated under the RJMCMC and MCMC approach, respectively. The coverage of the 95% credibility intervals is close to the nominal level. These results indicate that similar results can be achieved considering k unknown and fixing it on the true value. Although the MCMC algorithm has certain advantages with respect to the computational cost when compared to the RJMCMC, the number of components is generally unknown and estimate it can be a practical interest in the problem, thus the RJMCMC is a reasonable alternative to sample from the posterior distribution in this case.

Table 2: *Summary measurements for the point and 95% credibility interval estimates of the model parameters over 1000 simulations considering k unknown (RJMCMC) and known (MCMC).*

	μ_1	μ_2	μ_3	σ_1^2	σ_2^2	σ_3^2	w_1	w_2	w_3
RJMCMC									
SRMSE	0.10	0.48	0.45	0.11	0.81	0.38	0.05	0.05	0.08
MAE	0.08	0.36	0.22	0.07	0.63	0.18	0.05	0.04	0.05
Cov. (%)	96.7	94.1	90.0	92.0	96.2	97.6	100.0	99.9	90.6
Wid.	0.45	1.96	0.84	0.39	4.08	1.23	0.16	0.22	0.25
MCMC									
SRMSE	0.10	0.48	0.14	0.11	0.96	0.18	0.05	0.06	0.03
MAE	0.08	0.36	0.11	0.07	0.70	0.15	0.05	0.04	0.03
Cov. (%)	96.9	94.0	96.1	92.0	95.8	96.9	100.0	98.9	99.9
Wid.	0.45	1.96	0.59	0.40	4.19	0.81	0.16	0.23	0.25

4 Application to a real data set

We applied the methodology on a real data set that concerns antibody levels of Cytomegalovirus (CMV) in 5,126 individuals, both males and females, from 6 years to 49 years old. This data set was extracted from the 2003 - 2004 National Health and Nutrition Examination Survey (NHANES)¹.

The CMV is a member of the *Herpesviridae* family of viruses and, according to [5], is a common virus that occurs widely throughout the population but rarely causes noticeable symptoms or significant health problems.

One method of detecting a CMV infection is doing the antibody testing on blood samples. It can also be used to determine if someone has had recent or past exposure. There are two types of CMV antibodies that are produced in response to a CMV infection, IgM and IgG, and one or both may be detected in the blood. IgM antibodies are the first to be produced by the body in response to a CMV infection and they are present in most individuals within a week or two after the initial exposure. On the other hand, IgG antibodies are produced by the body several weeks after the initial CMV infection and provide protection from primary infections. Levels of IgG rise during the active infection, then stabilize as the CMV infection resolves and the virus becomes inactive. After a person has been exposed to CMV, the person will have some measurable amount of CMV IgG antibody in their blood for the rest of their life. CMV IgG antibody

¹Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2003 - 2004][<http://www.cdc.gov/nchs/nhanes>].

testing can be used, along with IgM testing, to help confirm the presence of a recent or previous CMV infection. Particularly, this data set consists of the IgG levels of CMV.

The range of values for the antibody levels CMV IgG are from 0.048 to 3.001. To the values reported as “out of range” (i.e. over the detectable range, > 3.00) the survey specialists usually assign the value of 3.001. Thus, there is a lot of individuals with this particular value on the data set. Figure 7 shows the antibody levels of CMV IgG distribution for 5,126 individuals infected and not infected. The interest here is in identifying subgroups of IgG as a marker of the presence of the disease.

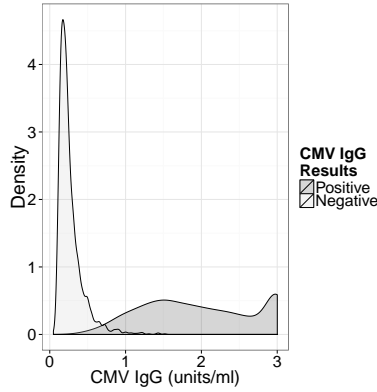


Figure 7: *Distribution of antibody levels of CMV IgG (units/ml) for 5,126 individuals infected by the virus or not.*

As shown in Figure 7, we clearly identify two or three heterogeneous subpopulations, so it is reasonable to fit a Normal mixture model on this data. On the inference we considered also the one estimating k and fixing it on its true value. For the RJMCMC simulations, we generated 45,000 samples from the posterior distribution, discarded the first 5,000, then thinned the chain by taking every 5th sample value. For the MCMC simulations, we considered 50,000 sweeps, then discarded the first 10,000 and thinned the chain by taking every 10th sample value.

Figure 8 displays the posterior distribution of k and predictive densities of antibody levels estimating k (RJMCMC) and fixing $k = 3$ (MCMC), represented by the dashed and dotted lines, respectively. The posterior distribution of k obtained from RJMCMC simulation favours 3 components, however, the third one is censored due to the group assigned as 3.001. It should be noted that predictive plots for RJMCMC and MCMC are very similar, showing a good performance even when k is estimated.

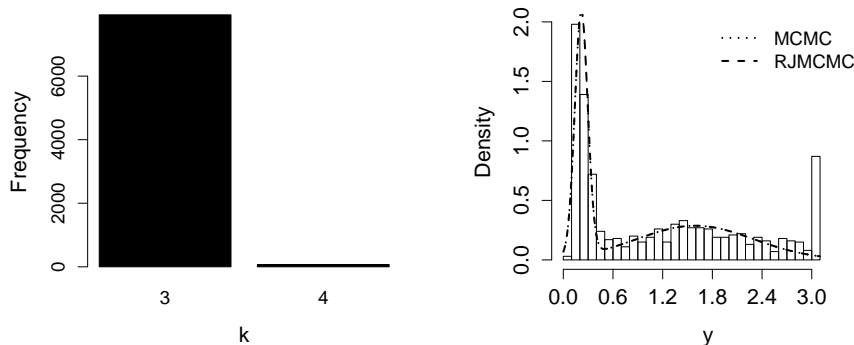


Figure 8: *Posterior distribution of k and predictive densities for the real data set.*

The model fit estimating k and fixing it in $k = 3$ results a DIC of -3692.59 and -3693.99 , respectively. Thus, as DIC increases with the number of parameters, it is expected that RJMCMC presents a higher DIC. However, since both DICs were very similar, it is possible to conclude that both methods are efficient in this case.

5 Conclusions and suggestions for future work

We have considered the problem of the fit of mixture models for heterogeneous populations under different levels of heterogeneity. We have discussed the improvement of the convergence when assigning a weakly informative prior distribution for the mixture proportions even for more homogeneous populations.

Finally, we have also evaluated the inference for the model when the number of components is unknown (RJMCMC) and when it is fixed in a known value (MCMC). We have concluded that when the number of mixture components is unknown, the RJMCMC is a feasible alternative, achieving similar results when this number is fixed in the true value. Nevertheless, it requires slightly bigger computational effort than MCMC. On the other hand, if we are not interesting in estimating this number, fixing it in a smaller value than the truth will generate poor estimates, however, similar results are obtained when fixing it in the true value or greater than this.

The main findings of this work encourage an extension of this study to other mixture distributions, as the Poisson model discussed in [11].

References

- [1] Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 363-375.
- [2] Green, P. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711-732.
- [3] Jasra, A., Holmes, C. and Stephens, D. (2005) Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 50-67.
- [4] Jordan, M. I. (2004) Graphical models. *Statistical Science*, 140-155.
- [5] Kusne, S., Shapiro, R. and Fung, J. (1999) Prevention and treatment of cytomegalovirus infection in organ transplant recipients. *Transplant infectious disease*, **1**, 187-203.
- [6] McLachlan, G. and Peel, D. (2004) *Finite mixture models*. John Wiley & Sons.
- [7] Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, **26**, 195-239.
- [8] Richardson, S. and Green, P. (1997) On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731-792.
- [9] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583-639.
- [10] Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 795-809.
- [11] Viallefont, V., Richardson, S. and Green, P. J. (2002) Bayesian analysis of poisson mixtures. *Journal of Nonparametric Statistics*, **14**, 181-202.